# 一覧性を考慮した動画要約法の構築

修士学位論文要約（令和３年３月）

# 一覧性を考慮した動画要約法の構築

北山　晃太郎
指導教員：乾　健太郎

# Creating Video Summarization Data forMultiple Key-frame Captioning Task

Kotaro KITAYAMA
Supervisor: Kentaro INUI

Massively large-scale videos have recently been available online. The automatic video summarization is one of the crucial technologies to alleviate the cost of developers and end-users to check the contents of videos. This paper specifically focuses on a video summarization task, which we call ”key-frame captioning”. This task requires the system to generate a predefined number of key-frame and description pairs that summarize the video well. We first develop a large-scaled key-frame captioning dataset by extending an existing dataset for video summarization since, unfortunately, no dataset has such annotations currently. And we proposes baseline experiment for the task in this paper. From the results of the experiment we gained new knowledge about the new task.

## 1. はじめに

近年，動画コンテンツは膨大な量があり，かつ，実時間を超えて増加してきている．各動画を試聴しながら所望とする動画か判断するのは膨大な時間がかかり非現実的であり，動画の内容を短時間で把握するシステムが求められる．そういったシステムはいくつも考えられるが，その中でも動画要約は有望なシステムの一つである．動画要約の従来研究は多く存在し，例えば動画から複数枚の重要な画像を抜き出す要約方法や，動画から重要なシーンをセグメントで切り出し，各セグメントに対して説明文を付与する要約方法などである．しかし，従来研究では「要約結果の提示方法」「短時間での動画の内容把握」という観点での明確な議論がない．本論文では，要約結果の一覧性という要素・観点に着目する．ここで，一覧性が高いとは短時間での内容把握が可能と仮定し，一覧性が高い動画要約という新タスク（キーフレーム説明文生成）を提案する．そのタスクでは，動画をいくつかの事前に固定された数（2 から 5 程度）の重要画像と説明文のペアで表現する．また，キーフレーム検出を行うための評価用データセットの作成とベースライン法の設計も同時に行う．

## 2. キーフレーム説明文生成タスクの設計

本研究では，新しい動画要約タスクとしてキーフレーム説明文生成を提案している．図 1 に示すように，同タスクでは動画を重要画像（キーフレーム）と説明文のペアの形式で要約する．ここで，正解のペアの数は予め与えられるという仮定のもと以降の実験を行った．

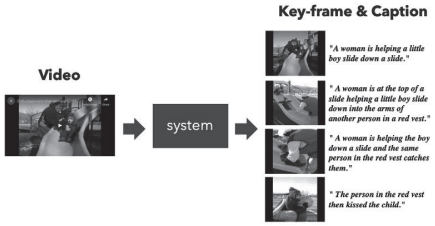


図 1 キーフレーム説明文生成の概要

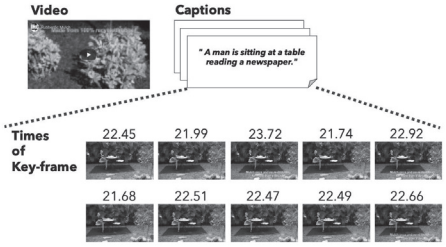


図 2 作成したデータセットの概要

要約を行うシステムを評価する際に，評価対象はキーフレームと説明文の2種類存在する．まず，キーフレームに関してはシステムが予測したキーフレームが正解データに含まれていた割合で評価する．また，フレーム単位で正解を当てることは困難だと考えられるため，予測結果と正解の間のコサイン類似度も同時に測定する．次に，説明文に関しては予測結果と正解の間の文としての類似度を測るために BLEU スコアを用いた．

## 3. データセットの構築

提案したタスクを解くためには，動画内の重要なイベントに対するキーフレームと説明文の情報を含んだデータセットを構築する必要がある．本論文では，動画説明文生成タスクのために作成された ActivityNet Captions データセットを拡張することによって所望とするデータセットの構築を試みた．ActivityNet Captions には動画内の重要なイベントに関して，そのセグメントと説明文のペアの情報が付与されている．説明文に該当するフレームをセグメントの中から選ぶことで追加のアノテーションを付与することによりデータセットの拡張を行った．また，データの作成には Yahoo!クラウドソーシングを用いた．以下の図 3 にアノテーションに用いたツールを示す．クラウドソーシングに参加したあのテーターは，動画を再生しながら，説明文に該当するフレームが表示されたタイミングでボタンを押すことによりアノテーションを付与することができる．


図 3 アノテーションツール

## 4. モデル

キーフレーム説明文生成を行うため，以下の図 4 に示すような2種類のモデルを組み合わせることにより，ベースラインとなるモデルを構築した．

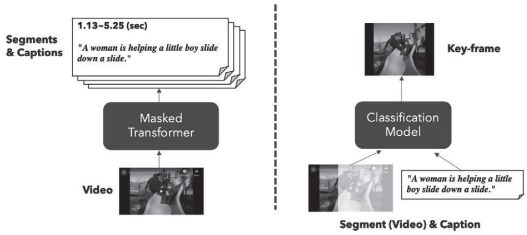


図 4 動画説明文生成モデル（左）と分類モデル（右）

2種類のモデルはそれぞれ，動画説明文生成で用いられているモデルとキーフレームのための分類モデルである．それぞれの役割としては，まず，動画説明文生成モデルは動画を入力として重要なシーンのセグメントと説明文のペアを出力する．また，分類器はそのペアの情報を入力とし，予測されたセグメントの中から説明文に該当するフレームの位置を推定する．そのようにすることにより，全体としては動画を入力として，動画内の重要なシーンに関する説明文とそれに該当するキーフレームのペアの出力を得るシステムを実現することができる．

## 5. 実験・結果

3章で作成したデータセットにデータフィルタリングを施し，残ったデータを用いてモデルを訓練し，実験を行った．モデルが予測したキーフレームが正解データに含まれていた割合は 0.103 であり，予測結果と正解の間のコサイン類似度は 0.839 であった．また，説明文に関する BLEU スコアは以下の表 1 のようになった．現状のスコアはまだまだ低いと考えられ，結果の定量的・定性的分析を行うことによって，スコアを向上させるための改善案について考察した．

表 1 説明文に関する BLEU スコア

| BLEU@1 | BLEU@2 | BLEU@3 | BLEU@4 |
|---|---|---|---|
| 18.8 | 3.4 | 1.0 | 0.4 |

## 6. まとめ

新しい動画要約タスクであるキーフレーム検出を提案し，そのためのデータセット作成とベースラインモデルの構築を行った．作成したデータで訓練したモデルを用いてキーフレーム検出タスクを解くことによって，現状のベースラインの性能の測定と，結果から今後の精度向上のための考察を行った．

## 文献

1) Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In International Conference on Computer Vision (ICCV), 2017.
2) Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
3) Ke Wang, Mohit Bansal, and Jan-Michael Frahm. Retweet wars: Tweet popularity prediction via dynamic multimodal regression. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1842–1851, 2018.